Intelligent Optimization of Machine Learning Methods

Jonathan Gillett

July 2015

1 Project Definition

Machine Learning is a branch of Computer Science that has since evolved from the study of pattern recognition and computational theory and into a highly diverse and widely used area of research. This evolution has empowered society with the almost limitless amount of data that has become available[8, 9].

Machine Learning focuses on the creation of algorithms, often rooted in artificial intelligence methods; in order to learn from, model, and make predictions based on data.[4]. These algorithms frequently involve the analysis and understanding of large data sets, often referred to as ”Big Data”, which are seemingly insurmountable using traditional statistical methods[8]. However, with the use of Machine Learning methods, models are created based on the data, in order to make predictions or decisions, rather than following deterministic program instructions[3].

An essential step in the analysis and modelling process is dimensionality reduction to reduce the negative affects of the “curse of dimensionality”. The “curse of dimensionality” often arises when an algorithm does not scale well to high-dimensional data, requiring time or memory that is exponential based on the dimensions of the data[5].

Dimensionality reduction, and by extension clustering, are quintessential operations in Machine Learning, with numerous applications in other fields. The most widely used methods in Machine Learning for clustering and dimensionality reduction are K-Means and Principal Component Analysis (PCA)[6, 5, 4].

In K-Means the data is clustered by first determining a set of centroid points. The algorithm then finds a partition, such that the squared error between the empirical mean of a cluster and the centroid points in the cluster is minimized[6, 3]. With PCA, the dimensionality of the data is reduced by projecting the N-dimensional data onto a k-dimensional linear subspace that minimizes the *reconstruction error*, the sum of the squared L_2-distances between the original and projected data. The resultant projection of this operation captures the most variance in the data within the orthogonal principal components of the linear subspace[7, 9, 3].

Therefore, the main focus of this research is to improve the adaptability of these widely used clustering and dimensionality reduction methods using intelligent methods in order to make them more robust to a wide variety of problems and to further enhance their accuracy and effectiveness.

2 Project Importance

Dimensionality reduction is essential in Machine Learning to reduce the negative affects of the "curse of dimensionality" and to simplify perform clustering to determine trends in the data are critical. Furthermore, clustering methods are critical to Machine Learning in order to find potential patterns and clusters of similar data, which are required in order to analyze the data and create models.

These widely used operations are often used to solve a wide variety of diverse problems in Machine Learning but lack the capability to intelligently adapt to the changing conditions of each and every problem. Furthermore, in particular for K-Means clustering, the operation is **NP-Hard**[6], making the problem computationally infeasible for highly dimensional or very large data sets, which is often the case when applying Machine Learning methods to Big Data.

Therefore, any improvements made to these widely used operations would have a profound impact on the accuracy and effectiveness of the methods. In addition, the enhancements provided by using intelligent soft computing techniques would make the methods more generalized and able to adapt to the diversity of data from different problem domains.

3 Approach

The approach will be to apply intelligent methods to Machine Learning clustering and dimensionality reduction methods to enhance their performance and adaptability. The Use of a single or combination of multiple intelligent methods (e.g. Fuzzy Logic, Evolutionary Computation, Neural Networks) will be applied and the results will be compared against the original clustering and dimensionality reduction methods for comparison.

The solution implementation will be written in C/C++ using modern software development practices such as Object Oriented Programming (OOP), C99[1] and C++11[2] standards, which adds many enhancements to C/C++ languages, and thorough documentation to ensure that the code is easily understood. In addition, a Makefile will also be provided, to simplify the process of building and running the executable.

4 Methodology

To successfully complete this project the methodology will consist of first determining a set of suitable test data, so that the benchmarking is consistent and produces accurate results. Following this, the current most-widely used algorithms for performing clustering and dimensionality reduction, K-Means and PCA, will be implemented. These implementations will then be executed with all of the test data to generate a basis of results. The enhanced versions of these algorithms will be implemented using intelligent methods. Lastly, these enhanced versions will be evaluated and compared to the base results of the original implementations.

4.1 Detailed Breakdown of Methodology

1. Establish a set of test data that is widely used for benchmarking and comparing clustering and dimensionality reduction algorithms in the Machine Learning community.

2. Implement the standard algorithms for clustering and dimensionality reduction, K-Means clustering and PCA.

3. Execute each of the algorithms with the test data set in order to establish a basis to compare to the enhanced implementations.

4. Implement the enhanced clustering and dimensionality reduction algorithms using a single or combination of multiple intelligent methods (e.g. Fuzzy logic, Evolutionary Computation, Neural Networks).

5. Execute the clustering and dimensionality reduction algorithms enhanced using intelligent methods with the test data in order to evaluate the solutions.

6. Compare the results of the enhanced algorithms to the standard algorithms, focusing on improvements in the accuracy and adaptability of the enhanced solution to different problem domains.

References

[1] ISO/IEC 9899 – Programming languages – C99 Standard. `http://www.open-std.org/JTC1/SC22/WG14/www/standards`, Dec. 1999. [Online; accessed Jul. 11, 2015].

[2] ISO/IEC 14882:2014(E) – Programming Languages – C++11 Standard. `https://isocpp.org/std/the-standard`, Jan. 2012. [Online; accessed Jul. 11, 2015].

[3] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[4] Thomas G Dietterich. Machine learning for sequential data: A review. In *Structural, syntactic, and statistical pattern recognition*, pages 15–30. Springer, 2002.

[5] Pedro Domingos. A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87, 2012.

[6] Anil K Jain. Data clustering : 50 years beyond K-means. *Pattern Recognition Letters*, 31(8):651–666, 2010.

[7] Ian Jolliffe. *Principal component analysis.* Wiley Online Library, 2002.

[8] Steve Lohr. The age of big data. *New York Times*, 11, 2012.

[9] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning.* MIT press, 2012.